DOCUMENT RESUME

ED 235 190 TM 830 580

AUTHOR Cook, Linda L.; And Others
TITLE An Assessment of the Relationship between the Assumption of Unidimensionality and the Quality of IRT True-Score Equating.
PUB DATE Apr 83
NOTE 69p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Equated Scores; Factor Analysis; *Latent Trait Theory; Testing Problems; True Scores
IDENTIFIERS College Board Achievement Tests; Scholastic Aptitude Test; *Unidimensionality (Tests); Violation of Assumptions

ABSTRACT

The purpose of this study was to empirically examine the relationship between violations of the assumption of unidimensionality, as assessed by the factor analysis of item parcel data, and the quality of item response theory (IRT) true-score equating, as measured by score scale stability. The verbal section of the Scholastic Aptitude Test (SAT) and the College Board Mathematics Level II examination were selected for use. Factor analyses were performed on each of the six selected test forms, using a correlation matrix of item parcel scores as input. The results of the factor analyses were related to the results of previous equating studies, hypothesizing that the equating chain that resulted in the least scale stability (SAT-verbal) would show evidence of greater multidimensionality than the equating chain (Mathematics Level II) that provided the superior equating results. The Mathematics Level II equating results were superior to the SAT-verbal equating results, and the dimensionality analyses revealed that the Mathematics Level II item parcels were more nearly unidimensional than the SAT-verbal item parcels. The dimensionality analyses also verified that SAT-verbal Form V4 and Mathematics Level II Form CC were each less parallel to the other two forms in their respective equating chains than the other forms were to each other. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED235190

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFOHMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

An Assessment of the Relationship Between the Assumption of Unidimensionality and the Quality of IRT True-Score Equating[1,2]

Linda L. Cook[3]
Neil J. Dorans
Daniel R. Eignor
Nancy S. Petersen

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. L. Cook
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

[1] A paper presented at the annual meeting of AERA, Montreal, 1983.

[2] The authors would like to acknowledge the assistance of Karen Carroll, Ted Blew, and Steve Dass in completing this study, and thank Karen Damiano for her patience and invaluable secretarial support.

[3] The authors names appear in alphabetical order.

TM 830 580

An Assessment of the Relationship Between the Assumption of Unidimensionality and the Quality of IRT True-Score Equating

Linda L. Cook
Neil J. Dorans
Daniel R. Eignor
Nancy S. Petersen
Educational Testing Service

INTRODUCTION

In recent years there has been considerable research and interest devoted to the use of item response theory (IRT) in the solutions to a variety of measurement problems (see Lord, 1980; Hambleton, 1983). Because of the special properties of test data characterized by IRT models, users are often able to solve problems not amenable to solution through the use of traditional psychometric methods. However, in order for IRT to be useful in the solution of measurement problems, certain fairly strong assumptions about the data must be met. One of the most important of these assumptions is the assumption of unidimensionality. Most IRT models that are currently used with binary scored item response data assume that the probability of a correct response to an item can be modeled by a mathematical function that assumes a single ability dimension is common to all items. For reasons to be developed later in this paper, researchers working with binary scored item response data typically assume that the items which appear to test a skill or content area are unidimensional (Divgi, 1981b). This assumption is almost surely inappropriate for many types of test data (Drasgow and Parsons, in press). The issue then becomes one of the consideration that even when an IRT model is not strictly appropriate for the data, it may still

be robust to violations of the assumption of unidimensionality for certain applications. The demonstration of the robustness of an IRT model to violation of the unidimensionality assumption for specific applications is clearly an empirical issue, though seldom are empirical studies of this sort seen in the literature. This lack of empirical verification is not caused by problems in the use of IRT methods in the particular application area as much as it is caused by the great difficulties involved in the assessment of the dimensionality of binary scored item response data.

A variety of methods have been advanced to date for assessing the unidimensionality assumption for binary scored item response data. If the one-parameter logistic model and conditional maximum likelihood estimation techniques are used, a number of statistical tests of the unidimensionality assumption follow directly from the estimation of item parameters over different groups of people or subsets of items (see Gustafsson, 1980; van den Wollenberg, 1982a, 1982b). If the one- or two-parameter normal ogive model and marginal maximum likelihood estimation procedures are used (Bock and Lieberman, 1970), a data-based test of the unidimensionality assumption can be developed. McDonald (1981, 1982), while presenting IRT models that utilize marginal maximum likelihood estimation procedures as special cases of the random regressors factor analytic model, has suggested that the set of residual item covariances after fitting a one factor model be studied for indications of departures from unidimensionality. Hattie (1981), in a large scale simulation study, studied McDonald's suggested procedure with a number of other proposed measures of unidimensionality and found

McDonald's suggestion provided the best results. Because the one-parameter or Rasch model is for the most part inappropriate for the analysis of binary scored multiple choice item response data (Fischer, 1978; Divgi, 1981a) and because researchers object to the assumption of normally distributed abilities, needed in the random regressors factor analysis model (McDonald, 1982), many researchers at present work with the three-parameter logistic model and unconditional maximum likelihood estimation procedures, as used, for instance, in the computer program LOGIST (Wingersky, Barton, and Lord, 1982). (See Bock and Aitken, 1981, however, for an approach that does not depend on the assumption of normally distributed abilities.) For this model and estimation procedure, direct statistical or data-based tests of the unidimensionality assumption do not (at present) follow directly from the parameter estimation process. Bejar (1980) has developed a procedure for assessing dimensionality that works well in this context, but the procedure requires apriori knowledge about the test items so that a subset of the total set of items can be formed that is clearly unidimensional. Because tnis information is usually unavailable, researchers working with multiple choice items have instead chosen to use (linear) factor analysis with individual item data to assess unidimensionality, usually working with phi, or when possible, tetrachoric correlation coefficients. The theoretical problems involved with using such a procedure with phi or tetrachoric correlations have been clearly pointed out by McDonald (1981) and the practical problems by McDonald (1967), McDonald and Ahlawat (1974), Hambleton and Rovinelli (1983), and Lord and Novick (1968, p.349). Basically, the problem can

be summarized as follows. If a linear factor analysis of item data is undertaken, using either phi or tetrachoric correlation coefficients, then artifactual factors may appear in the factor solution due to the non-linear relationship between the observed response data and the underlying trait (McDonald and Ahlawat, 1974). Further, as mentioned earlier, McDonald (1982) has pointed out that item response theory models are special cases of non-linear factor analytic models. If, in effect, a non-linear factor analytic model is necessary to characterize the relationship between the response to an individual item and the underlying trait or factor that the item measures, then any attempt to use a more simplistic linear factor analytic model, or indices based on that model, to assess unidimensionality is bound to be problematic. McDonald (1981) makes the following point concerning the use of indices based upon linear factor analysis of binary scored item data:

> Commonly the proportion of variance due to the first principal component is recommended as a decision criterion for unidimensionality, presumably because it is a crude indicator, in general an overestimate, of the proportion of variance due to the first common factor. However, it is important to recognize that there is no direct relationship between the proportion of variance due to the first common factor and the presence or absence of additional common factors.

Given the issues involved in the use of linear factor analysis with binary scored item response data, there appears to be two possible approaches to the problem of assessing unidimensionality for those models (and estimation procedures) where a clearly developed procedure is not at present available. Hambleton and Rovinelli (1983) have offered one possible approach to the problem, which is based on McDonald's (1981) suggested procedure for studying dimensionality with

the random regressors factor analysis model. This involves looking at the residual covariances between items after fitting a (non-linear) single factor model. An alternative procedure involves the use of item parcels, or mini-tests, made up of small collections of non over-lapping items thought to measure the underlying dimension or dimensions. Data on individual items are no longer used; some justification for aggregating the data into mini-tests comes from the summary section of McDonald's 1981 article:

> (1) In principle, a set of n tests or n binary items is unidimensional if and only if the set fits a (generally non-linear) common factor model with just one common factor.
>
> (2) In checking the unidimensionality of a set of tests, a simple, appropriate, ancillary assumption is that the regressions of the tests on the factors are linear.

If item parcel data is to be used in a factor analytic study, of serious concern is the method chosen for defining the subsets from the total set of items and then placing items into parcels within a subset. Cattell and Burdsal (1975) recommend doing two factor analyses, one on the items to define the item dimensions for forming subsets within which the parcels will be formed and then one on the parcels to assess dimensionality. Because the first factor analysis suggested involves all the problems inherent in the factor analysis of item data, it would appear that a non-factor analytic procedure for the formation of item subsets, such as using item types as defined by content specifications, is necessary. Another concern when using item parcel data in factor analytic studies is the unwanted propagation of difficulty factors (see Swinton and Powers, 1980). While the use of item parcel data instead of

individual item data in a factor analytic study may tend to "linearize" the basic non-linear relationship between observed response and underlying trait, and hence minimize the incidence of artifactual factors due to non-linearity (McDonald and Ahlawat, 1974), if the parcels are of differing difficulty, artifactual difficulty factors may result. These factors will inhibit a reasonable assessment of the dimensionality of the data.

PROBLEM AND PURPOSE

One application area in which a number of researchers have recently taken increased interest is the use of item response theory for score equating purposes (see Cook and Eignor, 1983). This increased interest is reflected in the number of large scale testing programs that are either using IRT equating or considering its use for operational score reporting purposes. For example, Educational Testing Service now uses IRT to equate the Scholastic Aptitude Test (SAT) (Petersen, Cook, and Marco, 1982), the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT), and the Test of English as a Foreign Language (TOEFL). As with many other applications of IRT, it has been assumed that either the test data being used in the equating process is unidimensional or that the IRT model, when used in the equating process, is sufficiently robust with respect to violations of unidimensionality. The latter assumption is one commonly shared, without empirical verification, by a number of researchers. Divgi (1981b) points out:

> Similarly, the effect of a given departure from model assumptions is likely to depend on whether the model is used to make predictions about single items as in tailored testing or bias analysis, or to deal with entire tests as in equating. Applications of the latter kind are more likely to be robust.

Clearly, empirical research on the robustness of IRT models to violations of the assumption of unidimensionality for equating applications in a variety of testing contexts is needed.

The purpose of this study was to empirically examine the relationship between violations of the assumption of unidimensionality, as assessed by the factor analysis of item parcel data, and the quality of IRT true-score equating, as measured by score scale stability.

OVERVIEW OF STUDY

Two examinations were selected for use in this study. These examinations are the verbal section of the Scholastic Aptitude Test (SAT) and the Mathematics Level II examination, both administered by Educational Testing Service for the College Board Admissions Testing Program. Both examinations have recently been used in studies of the assessment of scale stability resulting from the use of IRT true-score equating procedures; the results for SAT-verbal are presented in Petersen, Cook, and Stocking (in press) and the results for Mathematics Level II in Cook and Eignor (1983).

The two examinations used in this study were chosen for several reasons. First, they represent different content areas as well as different types of tests. The verbal section of the SAT is generally considered to be an aptitude test, i.e., it is designed to measure overall verbal ability. On the other hand, the Mathematics Level II test is an achievement test that is designed to measure specific content areas such as algebra and geometry. Secondly, the results of Petersen, et al, (in press) indicated that application of IRT equating methods

resulted in considerably less scale stability for the verbal section of the SAT than for the mathematical section. In contrast, the results of the Cook and Eignor (1983) study indicated that application of IRT equating methods to the Mathematics Level II test resulted in a high degree of stability in the equated scores.

As mentioned previously, both the Petersen, et al, (in press) study and the Cook and Eignor (1983) study used scale stability as a criterion for evaluating the equating results. Scale stability refers to the extent to which a scale maintains the same meaning over time, and can be assessed by equating a test form to itself through an intervening chain of test forms. The equating results used for the present study are based on a chain of six SAT-verbal forms and seven Mathematics Level II forms. For the factor analytic portion of the study, an attempt was made to isolate, within each equating chain, that pair of adjacent forms that appeared to be the least parallel. These two forms, as well as a form adjacent to one of the forms, were then selected for further study using factor analytic techniques.

Factor analyses were performed on each of the six selected test forms (three SAT-verbal forms and three Mathematics Level II forms). A correlation matrix of item parcel scores was used as input to the factor analyses. For the SAT-verbal forms, items were grouped into parcels based on four item types: sentence completions; antonyms; analogies; and items based on reading passages. Item parcels for the Mathematics Level II forms were constructed using five content subclassifications contained in the specifications for the test: algebra, geometry, trigonometry, mathematical functions, and a somewhat

more general subclassification related to number theory, logic and proof, and probability.

For each test form, a series of confirmatory factor analyses using the LISREL V computer program (Joreskog and Sorbom, 1981) were performed. Several factor analytic models were used, including a second order factor model, which is a special case of hierarchical factor analytic models (Schmid and Leiman, 1957). Drasgow and Parsons (in press) have used a second order factor model in their work involving the application of three-parameter model unconditional maximum likelihood estimation techniques, as operationalized by LOGIST (Wingersky, et al, 1982), to multidimensional data. An attempt was made to relate the results of the factor analyses to the results of the equating studies. It was hypothesized that the equating chain that resulted in the least scale stability (SAT-verbal) would show evidence of greater multidimensionality or lack of form to form parallelism than the equating chain (Mathematics Level II) that provided the superior equating results.

METHODOLOGY

Description of Tests

As mentioned in the previous section, two examinations were selected for this study. These examinations are the verbal section of the Scholastic Aptitude Test (SAT) and the Mathematics Level II examination. The verbal section of the SAT is a multiple choice test that has been described as measuring developed verbal reasoning abilities that are related to successful performance in college. It is intended to

supplement the secondary school record and other information about the student in assessing readiness for college-level work. The Mathematics Level II examination is a multiple choice achievement test that is used in conjunction with measures of high school performance, as well as other standardized tests such as the SAT, by colleges and universities in selecting students for admission and/or course placement.

Test specifications for SAT-verbal have not remained constant over years. Test booklets containing SAT forms administered prior to the Fall of 1974 consist of two 45-minute sections (one SAT-verbal and one SAT-mathematical) and three 30-minute sections (one SAT-verbal, one SAT-mathematical, and one experimental containing an anchor test or pretest). The two SAT-verbal sections contain a total of 90 five-choice items composed of 43 reading comprehension items (18 sentence completions and 7 reading passages each of which is followed by 5 items based on the passage) and 37 vocabulary items (18 antonym items and 19 analogy items). Of the SAT-verbal forms used in this study, only the form designated V4 was developed to these specifications. Test booklets containing SAT forms administered since the Fall of 1974, which includes the other SAT-verbal forms used in this study, consist of six 30-minute sections: two SAT-verbal sections, two SAT-mathematical sections, one Test of Standard Written English, and one experimental section. The two SAT-verbal sections contain a total of 85 five-choice items composed of 40 reading comprehension items (15 sentence completions and five reading passages each of which is followed by 5 items based on the passage) and 45 vocabulary items (25 antonym items and 20 analogy items).

All of the Mathematics Level II forms used in this study were developed from the same set of content specifications. Each form contains 50 five-choice items and is administered in a 60-minute time period. The test is composed of approximately equal parts of algebra, geometry, trigonometry, mathematical functions, and a more general subcategory consisting of such topics as number theory, probability, and logic and proof. Unlike the situation with SAT-verbal, however, it is not a requirement that test forms developed with these content specifications contain exactly the same number of items measuring each content category.

Raw scores on the Mathematics Level II tests are typically transformed to scaled scores on a 200 to 800 scale, used for score reporting purposes, via linear equating methods. Prior to 1982, raw scores on SAT-verbal were typically transformed to another 200 to 800 scale, also used for score reporting purposes, via linear equating methods. Since January of 1982, IRT true-score equating has been used to place SAT-verbal forms on scale. Raw scores on both tests are obtained scores that have been corrected for guessing. Raw scores are computed by the formula R-W/k, where R is the number of correct responses, W is the number of incorrect responses, and (k+1) equals the number of answer choices per item.

Data Collection

Two samples were randomly selected for each test form used in the equating chains and the subsequent factor analyses (see Table 1). Whenever possible, samples for the experimental equatings were selected from the same population (test administration) used when the test form

Table 1

Raw Score[a] Summary Statistics for SAT-verbal and Mathematics Level II Samples

Form	Admin Date	N	Total Test		Anchor Test		Anchor Test/Total Test Correlation
			Mean	SD	Mean	SD	
SAT-verbal Samples							
V4	12/73	2665	35.04	16.37	14.01	8.54	.88
X2	4/75	2686	35.24	15.27	13.65	7.95	.86
X2	4/75	2562	34.42	15.31	16.74	8.07	.86
Y3	6/76	2578	34.48	16.34	16.14	8.41	.88
Y3	1/78	2549	31.37	15.86	14.36	8.17	.88
B3	5/79	2700	36.40	15.80	16.38	8.06	.88
B3	5/79	2665	35.90	15.24	15.04	8.01	.87
Y2	4/76	2879	34.16	14.84	15.08	8.19	.87
Y2	4/76	2774	33.57	14.50	16.02	7.44	.86
Z5	12/77	2853	30.73	15.61	14.43	7.69	.87
Z5	12/77	2814	31.13	15.91	13.76	7.83	.87
V4	12/73	2670	34.66	16.11	15.04	7.94	.86
Mathematics Level II Samples							
CC	12/80	2117	24.49	9.63	8.59	3.73	.90
WC	1/74	2160	22.84	10.71	7.86	4.07	.92
WC	4/76	1917	21.47	11.14	7.27	4.17	.92
AC	12/78	2209	25.15	10.09	8.37	3.74	.91
AC	1/80	2343	24.56	10.42	7.69	3.59	.91
VC	1/73	2406	23.61	11.09	7.72	3.72	.92
VC	1/73	2406	23.61	11.09	9.96	4.59	.93
XC	1/75	2045	23.75	10.57	10.03	4.67	.93
XC	1/76	2025	24.04	10.60	9.70	4.29	.93
ZC	12/77	2081	23.82	9.64	9.91	3.88	.91
ZC	1/79	2600	22.92	10.27	9.22	4.57	.93
BC	12/79	2278	25.35	9.23	9.83	4.23	.92
BC	12/79	2278	25.35	9.23	8.73	3.40	.90
CC	12/80	2117	24.49	9.63	8.63	3.58	.90

[a]Raw scores are obtained scores that have been corrected for guessing.

was originally introduced and placed on scale. Table 1 contains descriptive information regarding the samples. The table includes raw-score summary statistics for the total test and anchor test (common items) as well as dates of the test administrations from which the samples were selected. It should be noted that the common items linking the adjacent SAT-verbal forms are external to these forms, i.e., the common items are contained in a separately timed section and do not contribute to the total verbal score. The common items linking adjacent forms of the Mathematics Level II test are internal common items, i.e. these items are imbedded in the respective test forms and do contribute to the total test score.

Equating Methodology

Study Design and Criterion for Evaluation

A problem related to evaluation of the results of any equating method concerns the choice of a criterion measure. Since it is usually impossible to determine what the true equating should be, i.e., the true criterion against which to judge the actual equating, other criterion measures, varying in degree of complexity and assumptions made, have often been devised. (See Cook and Eignor, 1983, for a review of some of the more commonly used criteria for equating studies.) The criterion used in the present study to evaluate the quality of the equatings was scale drift.

Scale drift is said to have occurred if the results of equating test form D directly to test form A is not the same as that obtained by equating test form D to test form A through intervening forms B and C.

In order to evaluate scale drift for the verbal section of the SAT and the Mathematics Level II examination, a closed circular chain of equatings was performed for each of the tests. Figure 1 contains a diagram of the two equating chains. Upper case letter and number combinations indicate particular test forms and the abbreviation CI indicates common items linking adjacent test forms. It is possible to use the equating chains shown in Figure 1 to equate a test form to itself through a number of intervening test forms. If no scale drift has occurred, the initial (criterion) and final scaled scores for the forms should be identical. Any discrepancy between initial and final scores for a test form is attributed to scale drift resulting from application of the particular equating method. The results of the IRT equatings were evaluated both graphically and analytically.

IRT Model and Parameter Estimation

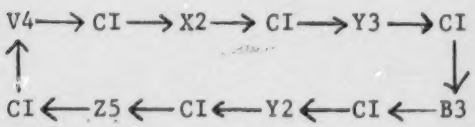
Item response theory (IRT) assumes that there is a mathematical function which relates the probability of a correct response on an item to an examinee's ability. (See Lord, 1980, for a detailed discussion.) Many different mathematical models of this functional relationship are possible. The model chosen for this study was the three-parameter logistic model.

The item parameters and examinee abilities for this study were estimated (calibrated) using the program LOGIST (Wingersky, et al, 1982; Wingersky, 1983). The estimates are obtained by a (modified) maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974).

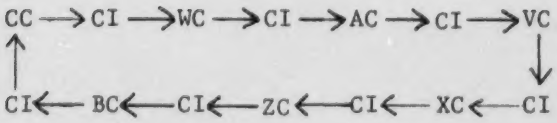
Figure 1

SAT-verbal and Mathematics Level II Equating Chains[a]

SAT-verbal



Mathematics Level II



[a]Letter and letter-number combinations indicate test forms. The abbreviation CI is used to indicate common items shared by two test forms.

LOGIST produces as output estimates of the item difficulty, item discrimination, and pseudo-guessing parameters for each item, and an ability (θ) parameter for each examinee. The metric chosen arbitrarily for the θ (and difficulty) scale is such that the distribution of estimates of θ has mean zero and standard deviation one. If two separate LOGIST runs are made for the same items, but different groups of examinees, the resulting parameter estimates will be on different scales.

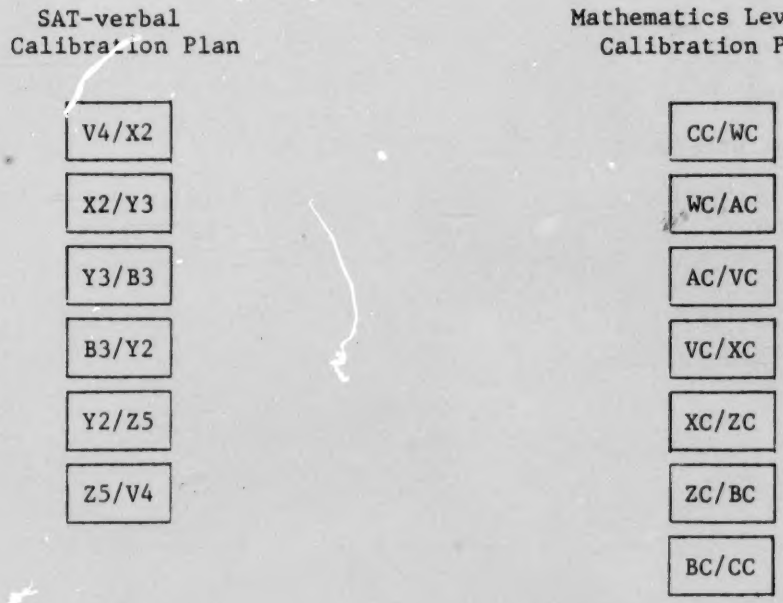
IRT Equating Method

The IRT equating method used in this study is referred to as IRT concurrent equating. (See Petersen, et al, in press, and Cook and Eignor, 1983, for detailed discussions of several IRT equating methods.) For IRT concurrent equating, each successive pair of test forms (e.g. SAT-verbal Forms V4 and X2) is calibrated in a single LOGIST run (see Figure 2). This results in item parameters on a common scale for each pair and allows direct equating of the two forms.

Once item parameter estimates on a common scale have been obtained, a number of different types of scores can be equated using item response theory; only true formula score equating was used for this study (Lord, 1980). The equating procedure was applied sequentially starting with the items calibrated in the first LOGIST run for each chain. Linear raw score to scaled score conversion parameters were already available to convert raw scores on each of the initial test forms in the two equating chains (i.e. SAT-verbal Form V4 and Mathematics Level II Form CC) to the 200 to 800 scales for these tests. As an example of the sequential equating process, consider the SAT-verbal equating chain. Equivalent

Figure 2

SAT-verbal and Mathematics Level II Calibration Plans[a]



[a]Boxes indicate separate calibration (LOGIST) runs. Each box represents a sample of approximately 4000 examinees (2000 examinees who took the new form of the test and 2000 examinees who took the old form of the test).

true formula score estimates were found for V4 and X2, resulting in a table of transformations of raw scores on X2 to the 200 to 800 scale. Form Y3 was then equated to X2, resulting in a table of transformations for raw scores on Y3 to the 200 to 800 scale. This procedure was repeated sequentially down both the SAT-verbal and the Mathematics Level II chains. The end product is a table of transformations of the raw scores on the initial form in each of the equating chains to the 200 to 800 scale.

Factor Analysis Methodology

Choice of Test Forms for Analysis

Only three test forms from each equating chain depicted in Figure 1 were chosen for the factor analyses performed for this study. The logic underlying the selection of the three forms was similar for both equating chains. An attempt was made to locate adjacent test forms that could be considered the least parallel and then to select a third form, adjacent to the pair, that could be considered reasonably parallel to the respective form, in the pair of forms, that it had been equated to. For the SAT-verbal chain, the obvious choice for the least parallel form in the equating chain was V4. As mentioned previously, this form contained five more items than any of the other forms in the chain and was built to different content specifications. The remaining two adjacent forms that were chosen were X2 and Y3. Both of those forms contained the same number of items, were built to the same content specifications, and were fairly similar both in reliability and overall difficulty level.

The choice of the three Mathematics Level II forms that were used for the factor analyses was not so straightforward. All of the forms in the Mathematics Level II equating chain were built to the same content specifications, contained the same total number of items, and were fairly similar in reliability and difficulty level. The three forms in the chain that were chosen were CC, WC and AC. The CC/WC pair was chosen because the equipercentile equating of the test forms that was carried out in the Cook and Eignor study (1983) indicated that the relationship between these forms was slightly curvilinear. Thus, it was concluded that of all of the pairs of test forms in the Mathematics Level II equating chain, CC and WC were the least parallel. Form AC was chosen because it was adjacent to WC. It should be emphasized that there was very little evidence of departures in parallelism for any of the test forms in the Mathematics Level II equating chain.

Formation of Items Parcels

Item parcel data were used in all the of factor analyses performed. Items from each SAT-verbal form were separated into item subsets on a within form basis using the four item types contained in the test: sentence completion items, antonym items, analogy items, and items based on reading passages. Within each of the four item subsets, items were placed into parcels of three to seven items each in a manner such that the mean difficulties of the parcels were approximately the same. The building of parcels of comparable difficulty was accomplished by assigning items to parcels based upon their equated delta difficulty indices. (See Hecht and Swineford, 1981, for an explanation of delta difficulty indices and the process of delta equating.) Within each of

the four subsets of items for SAT-verbal, the same number of parcels were formed across each of the three forms. Figure 3 contains the number of items within each of the four item subsets of SAT-verbal for each of the three forms and the number of parcels within each of the subsets.

Exactly the same procedure used for SAT-verbal was employed for forming the item parcels for Mathematics Level II except that the item subsets were formed using the five content subclassifications contained in the specifications for the test: algebra, geometry, trigonometry, mathematical functions, and the subclassification containing the areas of number theory, logic and proof, and probability. Figure 4 contains the number of items within each of the five item subsets of Mathematics Level II for each of the three forms as well as the number of parcels within each of the subsets.

Scores for examinees on the item parcels were formed, and then correlations were computed between parcels both within and across subtests for each form. The correlations among the parcels were used as input to the LISREL V program.

LISREL V: First-order and Second-order Models

The LISREL V computer program fits and tests models for linear structural relationships among quantitative variables. As mentioned earlier, the primary reason for developing item parcels was to yield variance-covariance matrices that were amenable to a linear factor analysis. Both first-order factor analysis and second-order factor analysis are special cases of the powerful LISREL V model. First-order factor analyses were employed in this study to assess the "effective"

Figure 3

Factor Pattern Matrices and Parcel Description for SAT-verbal Forms

$$A = \begin{bmatrix} X & 0 & 0 & 0 \\ X & 0 & 0 & 0 \\ X & 0 & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & 0 & X \\ 0 & 0 & 0 & X \\ 0 & 0 & 0 & X \\ 0 & 0 & 0 & X \\ 0 & 0 & 0 & X \end{bmatrix} \begin{matrix} \text{Parcels} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \\ 16 \\ 17 \end{matrix}$$

Parcels	Item Type	Number of Items	
		Form V4	Forms X2 and Y3
1-3	Sentence Completions	18	15
4-8	Antonyms	18	25
9-12	Analogies	19	20
13-17	Reading Passage items	35	25
	Totals	90	85

Figure 4

Factor Pattern Matrices and Parcel Description for Mathematics Level II Forms

$$A = \begin{bmatrix} X & 0 & 0 & 0 \\ X & 0 & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & 0 & X \\ 0 & 0 & 0 & X \\ X & X & X & X \\ X & X & X & X \end{bmatrix} \begin{matrix} \text{Parcels} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{matrix}$$

Parcels	Content Area	Number of Items		
		Form WC	Form AC	Form CC
1-2	Algebra	13	10	10
3-4	Geometry	9	10	9
5-6	Trigonometry	11	11	10
7-8	Functions	10	10	11
9-10	Miscellaneous	7	9	9
	Totals	50	50	50

dimensionality of the item parcels, i.e., the number of factors needed to adequately describe the covariation among item parcels. Second-order factor analyses were employed to test meaningful hypotheses about the structure of the data, hypotheses that were suspected to be pertinent to the quality of equating results.

LISREL V's Indices of Fit

LISREL V provides several indices of fit that are described by Joreskog and Sorbom (1981). When LISREL V provides maximum likelihood estimates of free parameters, it also provides the likelihood ratio χ^2 statistic with associated degrees of freedom and probability level. This index is most helpful in assessing competing models for the data because the difference in χ^2 values is itself distributed as a χ^2 with degrees of freedom equal to the difference in degrees of freedom associated with the two competing models. When one model is a special case of the other model, this difference in χ^2 values indicates whether the parameters that are estimated in the more general model add anything to the fit of the model for the data.

In addition to the likelihood ratio χ^2 statistic, LISREL V provides an adjusted (for degrees of freedom of the model) goodness of fit statistic, which for the maximum likelihood solution is

$$(1) \qquad GFI = 1 - \left[k(k+1)/2df \right] \left[\frac{\text{trace } (\hat{C}^{-1}C - I)^2}{\text{trace } (\hat{C}^{-1}C)^2} \right] ,$$

where C is the observed covariance matrix, $\hat{C}$ is the fitted covariance matrix, k is the number of observed variables, and df is the number of degrees of freedom. The GFI index, which typically ranges from zero to

one, is a measure of the proportion of covariation in the data accounted for by the model that produces $\hat{C}$.

Another overall goodness of fit index provided by LISREL V is the familiar root mean square residual,

$$(2) \qquad RMSR = \left[2 \sum_{i=1}^{n} \sum_{j=1}^{i} (c_{ij} - \hat{c}_{ij})^2 / k(k+1) \right]^{1/2} ,$$

where k is the number of observed variables, and c_{ij} and $\hat{c}_{ij}$ are elements of the observed and fitted covariance matrices. The RMSR index is useful for comparing the fit of two different models for the data.

In addition to these indices of global fit, LISREL V provides individual residuals in both raw and normalized forms. The raw residual is simply $c_{ij} - \hat{c}_{ij}$. The standardized residuals are taken from standard asymptotics based on normality, which states that the residuals have an asymptotic distribution with mean of zero and variance of $(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2/N)$, where N is the number of observations. Therefore, the standardized residual

$$(3) \qquad N^{1/2} (c_{ij} - \hat{c}_{ij}) / (\hat{c}_{ii}\hat{c}_{jj} + \hat{c}_{ij}^2)^{1/2}$$

is asymptotically a standard normal variable. Joreskog and Sorbom (1981) suggest that standardized residuals with values greater than two in absolute value merit close examination. For an effective summary of the fit of individual models, LISREL V presents Q-plots of the normalized residuals against normal quantiles. The slope of the plotted points are indicative of model fit. It is possible to evaluate model fit by visual inspection of the Q-plots. One can imagine a straight line passing through the plotted points and compare the slope of this

line with a 45 degree line represented on the plots by small dots. Slopes which are close to one represent moderate fit and those smaller than one poor fit. Perfect fit is represented by points falling in a straight line perpendicular to the abscissa.

First Order Common Factor Model

The traditional first-order common factor model is

(4) $$y = Ax + Du,$$

where

y is an n-by-1 vector of observable scores on the n item parcels,

x is a k-by-1 vector of non-observable scores on the k common factors that account for covariation among the n parcels,

A is an n-by-k matrix of common factor loadings or weights describing the regressions of the n parcel scores on the k factor scores,

u is an n-by-1 vector of unobservable unique scores, which could be further decomposed into measurement error and scores on specific factors, and

D is an n-by-n diagonal matrix of uniqueness loadings.

The n-by-n covariance matrix among the item parcels can be expressed as

(5) $$C_{yy} = AC_{xx}A' + D^2,$$

where

C_{xx} is the k-by-k matrix of factor covariances, and

D^2 is an n-by-n diagonal matrix of unique variances.

One goal of a factor analysis is to identify the number of common factors needed to fit the off-diagonal elements of C_{yy}. This is known

as the number of factors problem. First-order factor models, like that depicted in (4) and (5), were applied to the data to answer the number of factors question.

LISREL V was used to assess the number of factors problem in the following fashion. For each test form studied, the fit of a single common factor model to the correlation matrix among item parcels (correlation matrices were used to simplify proportion of variance interpretations and reduce the impact of variable length parcels on the multifactor solutions), was examined. Next, the fit of a very general two common factor model to the same data was examined. The two common factor models were essentially unconstrained in that no restrictions were imposed on the factor weight matrix A. Consequently, the two factor solutions were not readily interpretable. They did, however, permit assessment of the number of factors question.

Second Order Factor Model

To achieve interpretable results, a second-order factor model was used in a more classic confirmatory application of the LISREL approach. A second-order factor analysis can be thought of as a factor analysis of the first-order factors. It is a particularly fruitful approach to employ when one suspects that correlations among the first order factors can be explained by a single general factor. Such a model is particularly applicable to item data that one suspects is essentially unidimensional. Drasgow and Parsons (in press) suggested a second-order factor model that was influential in the selection of the approach used in this study to assess the dimensionality of item data.

The second-order factor model fitted to the first order common factors, x is

(6) $$x = bz + Fv,$$

where

z represents a score on the second-order general factor,

b is the k-by-1 vector of loadings of the k first order factors on z,

F is a k-by-k diagonal matrix of loadings of the k first-order factors on their corresponding group factors, and

v is a k-by-1 vector containing the k group factor scores.

This second-order factor model decomposes each first-order factor into a general factor that influences all first-order factors, and a group factor which influences performance only on that first-order factor. If the contribution of the general factor to every first order factor is large, the correlations among the first order factors will be close to unity. If the group factor for a particular first-order factor is relatively large, then the correlations of that first-order factor with other first-order factors will be among the lowest in the first-order factors correlation matrix.

As with the first-order factor analyses, the fit of the second-order factor models to the data was assessed. More importantly, substantive interpretations were attached to the second-order solutions. The substantive interpretations followed from the nature of the item parcels.

For the three SAT-verbal test forms, 17 parcels were constructed: three sentence completions parcels; five antonyms parcels; four analogy

parcels; and five parcels for items based on reading passages. The first-order factor weight matrix is highly restricted with simple structure corresponding to item type. In other words, the three sentence completions parcels load on a sentence completions factor only, the five antonyms parcels load on the antonyms factor only, etc. (See Figure 3 for a more detailed summary of the parcels and simple structure.) Thus, the second-order factor model contains a general verbal factor and four independent group factors corresponding to each of the four verbal item types. To the extent that the first-order factor variance explained by the general factor is large, the data is unidimensional. On the other hand, a sizeable group factor on a particular item type, say reading passage items, would indicate that this item type is making the largest contribution to violations of unidimensionality.

For the three Mathematics Level II test forms, 10 parcels were constructed: two algebra; two geometry; two trigonometry; two functions; and two miscellaneous (based on the general subcategory that included number theory, logic and proof, and probability). The factor weight matrix for these ten parcels is simple structure for the first eight parcels, i.e., the two algebra parcels load on an algebra factor only, the two geometry parcels on a geometry factor only and so forth. The last two rows of this 10-by-4 weight matrix contain free elements, which allows the miscellaneous item parcels to load on all four first-order factors. (See Figure 4 for a more detailed description.) The second-order factor model therefore contains a single general mathematics achievement factor and four independent group factors related to the four major content areas.

In sum, both first-order factor analyses and second-order factor analyses were employed. The first-order analyses focused on the number of factors or "effective" dimensionality issue. The second-order analyses were more confirmatory and focused on assessing hypothesized structures suggested by the item types and content areas measured by the tests. Fit of the model to the data was the dominant concern in the first-order analyses. Decomposition of first-order factor variance into a general and group specific component was the main concern of the second-order analyses. It was hypothesized that the stability of this decomposition across test forms is related to quality of equating.

RESULTS

IRT Equating

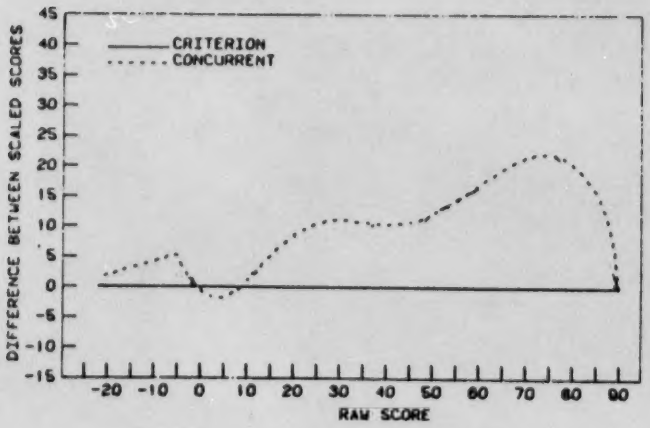
The final and initial (or criterion) conversions of SAT-verbal Form V4 and Mathematics Level II Form CC raw scores to their respective 200 to 800 scales should be identical. Departures resulting in scale drift may be due to sampling error and/or model fit problems.

To illustrate the extent to which the final and criterion conversions differ, scaled score differences (final minus criterion) for SAT-verbal and Mathematics Level II raw scores on the respective forms V4 and CC are shown graphically in Figure 5. The verbal scaled score discrepancies shown in Figure 5 indicate that the final conversion resulting from the IRT concurrent equating method overestimated the initial scale value for practically all of the raw score range. Examination of the Mathematics Level II scaled score discrepancies shown in Figure 5 indicates that the IRT concurrent method has a tendency to

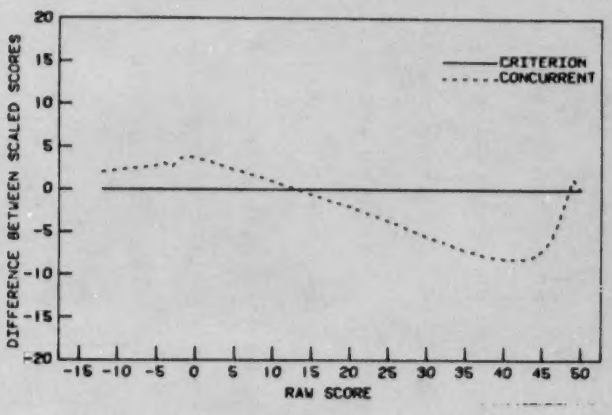
Figure 5

Summary of Equating Results for SAT-verbal and Mathematics Level II Equating

SAT-verbal



Mathematics Level II



Summary Statistics and Discrepancy Indices[1]

	SAT-verbal		Mathematics Level II	
	Form V4 Initial Scale (Criterion)	IRT Concurrent Equating	Form CC Initial Scale (Criterion)	IRT Concurrent Equating
Scaled Score:			Scaled Score:	
Mean	435.37	445.73	Mean	650.13	646.75
Standard Deviation	109.09	112.89	Standard Deviation	82.94	80.16
Total Error[2]		125.15	Total Error[2]		19.92
Bias		10.36	Bias		-3.39
S.D. of Difference		44.23	S.D. of Difference		2.91

[1]Computed for SAT-verbal raw scores 1 through 80 (N = 231,155) and for Mathematics Level II raw scores -2 through 49 (N = 14,744).

[2]Total Error = $(\text{SD of Difference})^2 + (\text{Bias})^2$

underestimate criterion scores for raw scores greater than 15 and to overestimate criterion scores for raw scores less than 15. It should be noted that application of IRT equating to the SAT-verbal chain resulted in a maximum scaled score discrepancy of close to 25 scaled score points, whereas the IRT concurrent method applied to the Mathematics Level II chain resulted in a maximum scaled score discrepancy of less than 10 scaled score points.

Observations based on the plots presented in Figure 5 are given more precise meaning by computing a discrepancy index for each comparison with the criterion. For each raw score x on the initial forms in the equating chains (SAT-verbal Form V4 and Mathematics Level II Form CC) there is a corresponding initial (criterion) scaled score t and an estimated scaled score t' derived from a specific equating method. The smaller the difference d between t and t', the smaller the scale drift and the more stable the equating method. A weighted mean square difference was used to summarize the differences between t and t'. The weighted mean square difference or total error is equal to the variance of the difference plus the squared bias, that is,

$$(7) \qquad \sum_j f_j d_j^2/n = \sum_j f_j (d_j - \bar{d})^2/n + \bar{d}^2, \text{ or}$$

(Total Error) = (Variance of Difference) + (Squared Bias)

where $d_j = (t'_j - t_j)$, t'_j is the estimated scaled score for raw score x_j, t_j is the initial or criterion scaled score for x_j, f_j is the frequency of x_j, $n = \sum_j f_j$, and $\bar{d} = \sum_j f_j d_j/n$. Summary statistics and discrepancy indices for each of the equating chains are also given in Figure 5. The values in Figure 5 were computed summing over SAT-verbal raw scores 1 to

80[1] and Mathematics Level II raw scores -2 to 49[1], using frequencies for the total group taking SAT-verbal Form V4 when it was first administered in December 1973 and Mathematics Level II Form CC when it was first administered in December 1980.

Examination of the verbal data presented in Figure 5 indicates that the IRT concurrent equating method overestimated both the mean and standard deviation of the criterion scaled scores. Bias accounted for approximately 86 percent of the total error. The information for Mathematics Level II summarized in Figure 5 indicates that the IRT concurrent equating method underestimated both the criterion mean and standard deviation. For the Mathematics Level II equating chain, bias accounted for approximately 58 percent of the total error.

Because of differences in test lengths and raw score frequencies of the groups used to weight the discrepancy indices, comparisons between the sizes of the total error for the two equating chains may be misleading. However, the discrepancy between this index for the two equating chains is so large that it would appear reasonable to conclude that the equating results for the Mathematics Level II chain are definitely superior to those for the SAT-verbal chain. Further evidence of the superiority of the Mathematics Level II results is provided by an examination of the scaled score means and standard deviations resulting from application of the IRT equating method to the two test chains. For the verbal chain, the IRT results overestimate the criterion mean by almost ten scaled score points and the criterion standard deviation by

---

[1]The discrepancy indices reported in this paper were computed as part of the Petersen, et al, (in press) and Cook and Eignor (1983) studies. For these studies, discrepancy indices were computed over the range of scores for which equipercentile raw to scaled score conversions were available. Had the total raw score range been included, changes in the discrepancy indices would have been negligible due to the low frequency of occurrence of scores in the extremes of the score scale.

approximately four scaled score points. On the other hand, for the Mathematics Level II chain, the IRT method underestimated both the criterion mean and standard deviation by approximately three scaled score points.

Factor Analyses

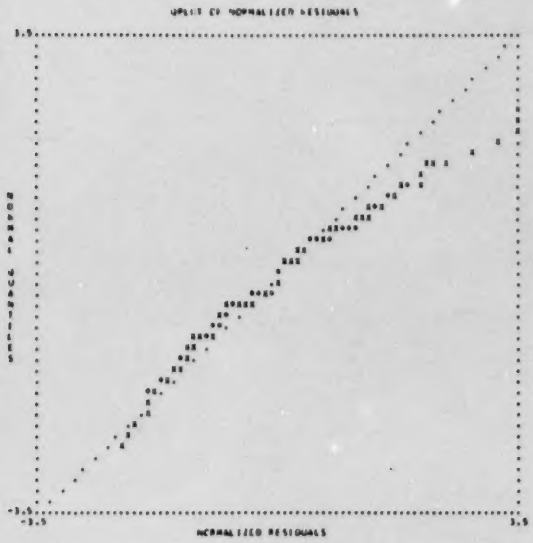
The factor analytic results are presented in the following fashion. The SAT-verbal results precede the Mathematics Level II results. For each test form, the number of factors question is assessed by examining the fit of first-order factor solutions. Then comparibility of the hypothesized second-order factor structures is examined across the three tests forms.

SAT-verbal

Number of factors. Figure 6 contains Q-plots of normalized residuals (see Methodology Section for detailed description of these plots) and indices of fit for SAT-verbal Form V4. There are four panels in this figure. The top two panels summarize the fit of a one factor first-order solution and a two factor first-order solution respectively, while the bottom two panels summarize the fit of two second-order factor solutions: a solution with one general second order factor and four group factors (one each for sentence completions, antonyms, analogies, and items based on reading passages), and a solution with two independent general factors and the same four group factors. The top left panel reveals that a single first-order factor solution does not fit the V4 item parcel correlation matrix. The residuals plot reveals a sizeable number of large positive residuals, which is indicative of underfactoring. In the top right panel it can be seen that adding a

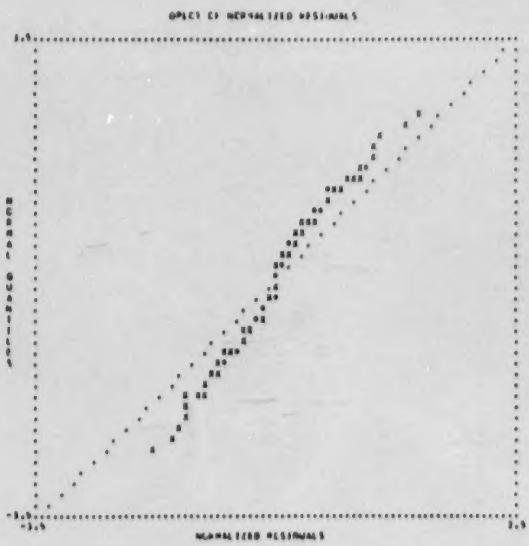
Figure 6

Normalized Residuals Plots and Indices of Fit for SAT-verbal Form V4



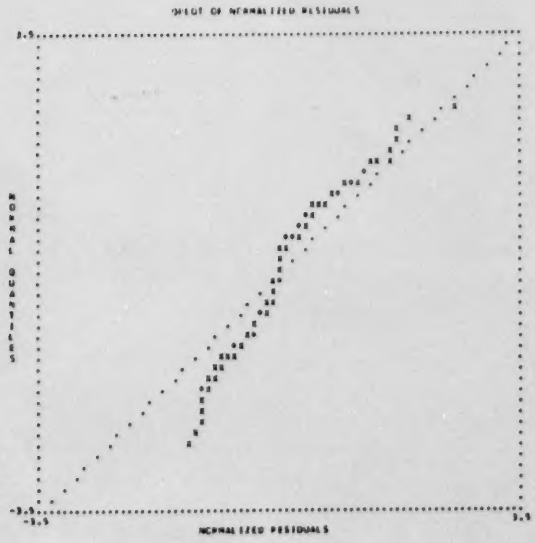
One Factor First-Order Solution

Chi Square = 604.55; df = 119
GFI = .958
RMSR = .026



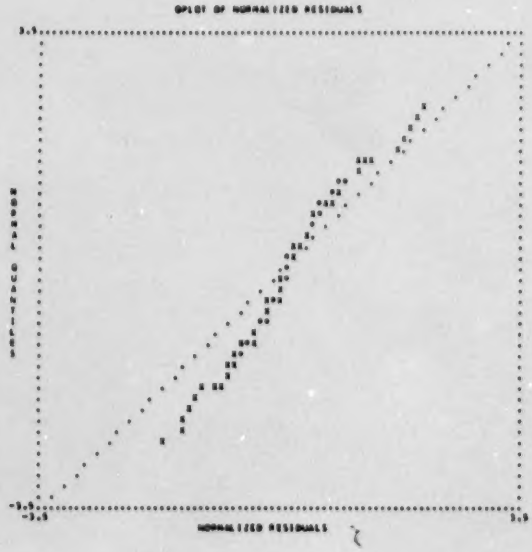
Two Factor First-Order Solution

Chi Square = 181.96; df = 101
GFI = .987
RMSR = .013



One General Factor and Four Group Factors Solution

Chi Square = 175.98; df = 115
GFI = .989
RMSR = .014



Two General Factors and Four Group Factors Solution

Chi Square = 151.66; df = 114
GFI = .990
RMSR = .013

second first order factor results in a very noticeable improvement in fit: The root mean square residual (RMSR) is halved from .026 to .013, the goodness of fit index (GFI) increases, and the chi square exhibits a sizeable drop from 604.55 (df=119) to 181.96 (df=101), an unquestionably significant improvement in fit.

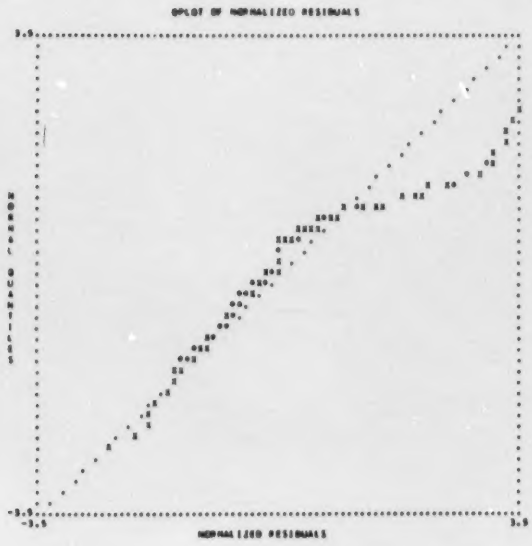
The information contained in the bottom left panel of Figure 6 reveals that a second-order solution with a restrictive factor pattern (see Figure 3), one general factor and four group factors, fits the V4 item parcel correlations very well. Adding a second general factor, orthogonal to the first (the bottom right panel in Figure 6), produces a slight but statistically significant improvement in fit, dropping the chi square from 175.98 (df=115) to 151.66 (df=114).

Figure 7 contains the normalized residuals plots and indices of fit for SAT-verbal Form X2. As was the case for Form V4, comparison of the top two panels reveals that one factor is clearly inadequate and addition of the second first order factor improves the fit noticeably. In fact, three first-order factors are really needed to provide a tight fit to the data. In order to verify this, the authors performed a three factor first order analysis (the results do not appear in Figure 7). Taking a third first-order factor results in a chi square of 124.29 (df=82), a GFI of .989, and RMSR of .010.

Contrast the fit portrayed in the bottom panels with the fit in the top panels. Fitting a restrictive confirmatory second-order solution that is theory-based fits better than the less restrictive first-order factor solutions. The lower left panel reveals that one general factor and four group factors fits the X2 item parcels correlation matrix very

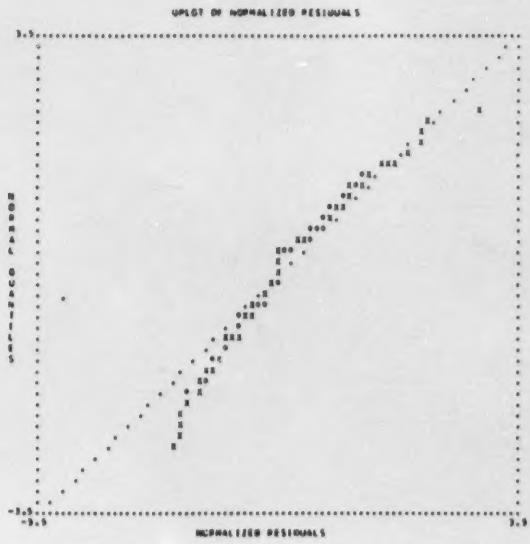
Figure 7

Normalized Residuals Plots and Indices of Fit for SAT-verbal Form X2



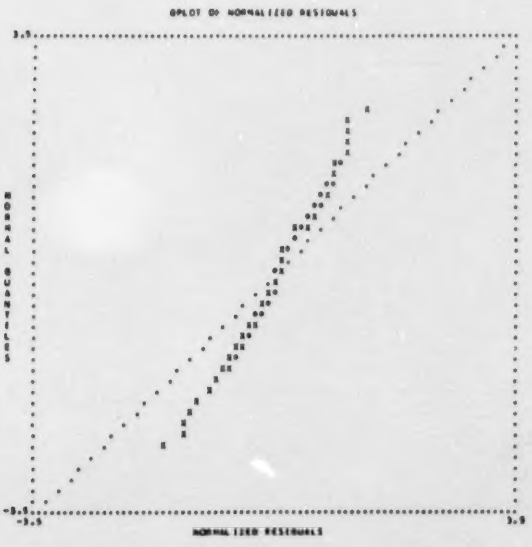
One Factor First Order Solution

Chi Square = 681.29; df = 119
GFI = .954
RMSR = .027



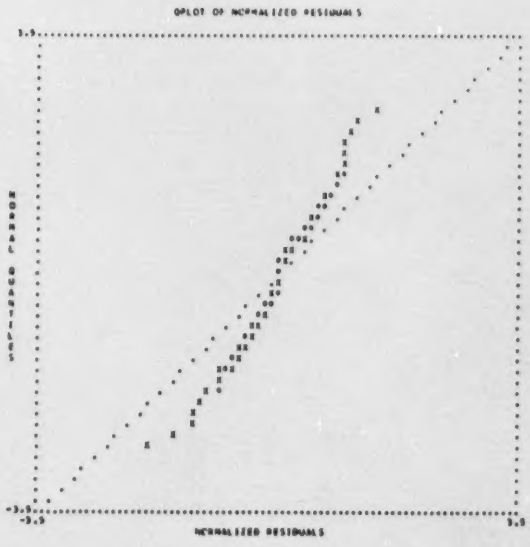
Two Factor First Order Solution

Chi Square = 309.76; df = 101
GFI = .976
RMSR = .017



One General Factor and Four Group Factors Solution

Chi Square = 145.11; df = 115
GFI = .991
RMSR = .012



Two General Factors and Four Group Factors Solution

Chi Square = 143.08; df = 114
GFI = .991
RMSR = .012

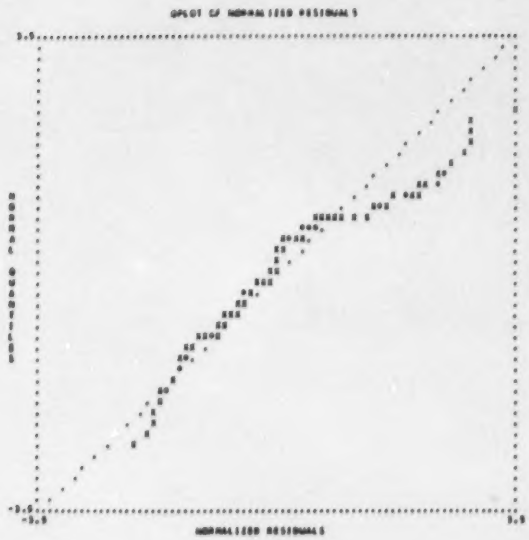
well. From the information displayed in the lower right panel, it can be seen that adding a second general factor is unnecessary. Thus a model that requires only one general factor to account for correlations between parcels composed of different item types fits the data very well. Recall, that for V4 the addition of a second general factor improved the fit slightly but significantly.

Figure 8 summarizes the fit results for SAT-verbal Form Y3. As was the case for Form X2, at least two first order factors are needed to fit the Y3 items parcels correlations. As with Form X2, the second-order solution with one general factor and four group factors provides a very good fit to the data. Adding a second general factor improves the fit very little.

Second-order structures. For all three SAT-verbal forms, the hypothesized second-order factor solutions fit the data well. Table 2 contains a numerical summary of the single general factor solutions (lower left panels in Figures 6-8). Here the relative contributions of the general factor and each the four group factors to the first-order parcel factors are tabled. In addition, Table 2 contains the correlations among the four first-order factors. One aspect of the data presented in Table 2 is immediately obvious. For every verbal form, the general factor is large relative to the group factors. This fact can be observed in the first-order factor correlations, all of which are .80 or higher, and in the variance contributions portion of the table. For example, for Form V4, the general factor accounts for 98 percent of the sentence completions factor variance, 85 percent of the antonyms factor variance, 93 percent of the analogies factor variance, and 82 percent of the reading passage items factor variance.

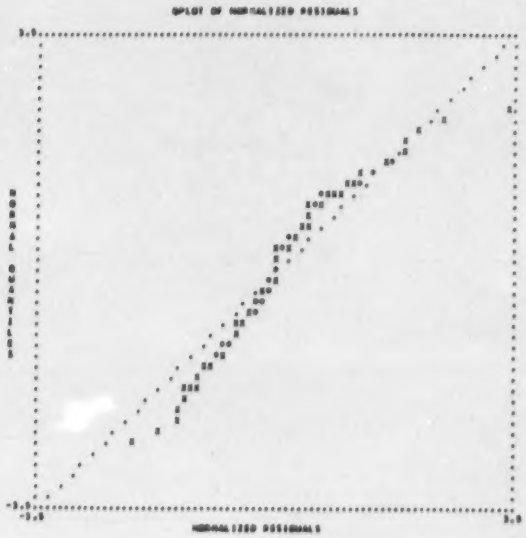
Figure 8

Normalized Residuals Plots and Indices of Fit for SAT-verbal Form Y3



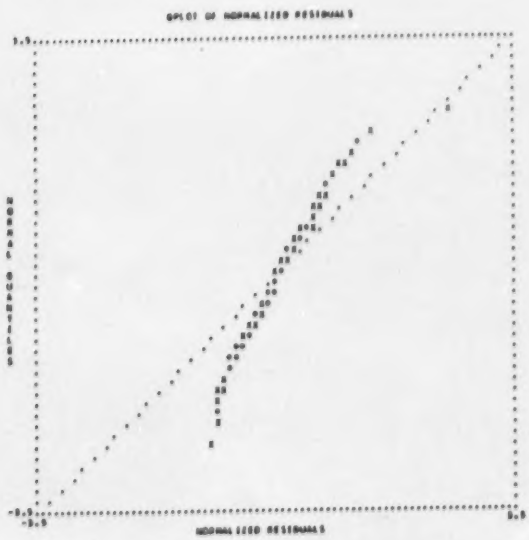
One Factor First Order Solution

Chi Square = 652.52; df = 119
GFI = .956
RMSR = .026



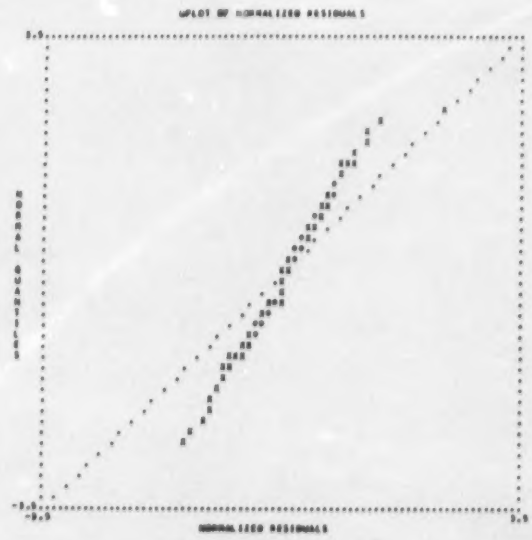
Two Factor First Order Solution

Chi Square = 296.06; df = 101
GFI = .978
RMSR = .017



One General Factor and Four Group Factors Solution

Chi Square = 169.12; df = 115
GFI = .989
RMSR = .013



Two General Factors and Four Group Factors Solution

Chi Square = 162.66; df = 114
GFI = .990
RMSR = .017

Table 2

Relative Contributions of One General and Four Group Factors to Variance of First Order Parcel Factors for Three SAT-verbal Forms

Test Form		First Order Factors				First Order Factor Correlations				
		Sentence Completions I	Antonyms II	Analogies III	Reading Passage Items IV					
							I	II	III	IV
	general factor	.98	.85	.93	.82	I	1.0			
						II	.92	1.0		
V4						III	.96	.89	1.0	
	group factors	.02[1]	.15	.07	.18	IV	.90	.84	.88	1.0
							I	II	III	IV
	general factor	.97	.92	.82	.81	I	1.0			
						II	.94	1.0		
X2						III	.89	.87	1.0	
	group factors	.03[1]	.08	.18	.19	IV	.89	.86	.81	1.0
							I	II	III	IV
	general factor	.96	.88	.86	.84	I	1.0			
						II	.92	1.0		
Y3						III	.91	.87	1.0	
	group factors	.04[1]	.12	.14	.16	IV	.90	.86	.85	1.0

[1]Not significantly different from zero ($p<.01$)

Looking across test forms (down columns in the table), it can be seen that the general factor accounts for almost all of the sentence completions factor variance on all three test forms. In contrast, the reading passage items factor has the largest group factor on all three forms. For Form V4, the general factor is more closely related to the analogies factor than the antonyms factor; for Form X2, the opposite is true. For Form Y3, the general factor is only slightly more related to the antonyms factor than it is to the anologies factor.

Figures 6-8 include a description of the fit of a second-order solution that allowed for a second general factor. Table 3 summarizes these solutions. It can be seen from the information summarized in Table 3, that for test Forms X2 and Y3, inclusion of a second general factor adds nothing to the solution. This fact can be observed in the miniscule contributions of this second general factor (.00 or .01) to first-order factor variance. Note also that for Forms X2 and Y3, the correlations among first-order factors remained virtually unchanged when the second general factor was added (compare correlations in Tables 2 and 3).

In contrast, addition of a second general factor has an impact on the solution for Form V4. Note that the antonym group factor is reduced substantially, while the reading passage item factor is reduced somewhat. This second general factor makes a non-trivial contribution to the variance of the antonym and reading passage item factors. As the footnote to the table indicates, this second general factor has positive weights for the vocabulary item types, antonyms and analogies, and negative loadings for the reading item types, sentence completions and

Table 3

Relative Contributions of Two General and Four Group Factors to Variance of First Order Parcel Factors for Three SAT-verbal Forms

Test Form		First Order Factors				First Order Factor Correlations				
		Sentence Completions I	Antonyms II	Analogies III	Reading Passage Items IV		I	II	III	IV
V4	general factor 1	.96	.91	.92	.84	I	1.0			
	general factor 2[1]	.00	.06	.00	.06	II	.92	1.0		
	group factors	.04	.03	.08	.10	III	.94	.92	1.0	
						IV	.91	.81	.87	1.0
							I	II	III	IV
X2	general factor 1	.97	.92	.82	.81	I	1.0			
	general factor 2[1]	.01	.00	.01	.01	II	.94	1.0		
	group factors	.02	.08	.17	.18	III	.89	.87	1.0	
						IV	.89	.86	.81	1.0
							I	II	III	IV
Y3	general factor 1	.96	.89	.86	.84	I	1.0			
	general factor 2[1]	.01	.01	.01	.01	II	.92	1.0		
	group factors	.03	.10	.13	.15	III	.90	.88	1.0	
						IV	.91	.86	.84	1.0

[1]For all three test forms, first order loadings on general factor 2 were positive for analogies and antonyms and negative for sentence completion and reading passage item parcels. With the exception of antonyms and reading passage items on Form V4, these loadings on the second general factor were trival.

reading passage items. Consequently, inclusion of the second general factor increases the correlations between the vocabulary item type factors, and decreases their correlations with the reading item type factors.

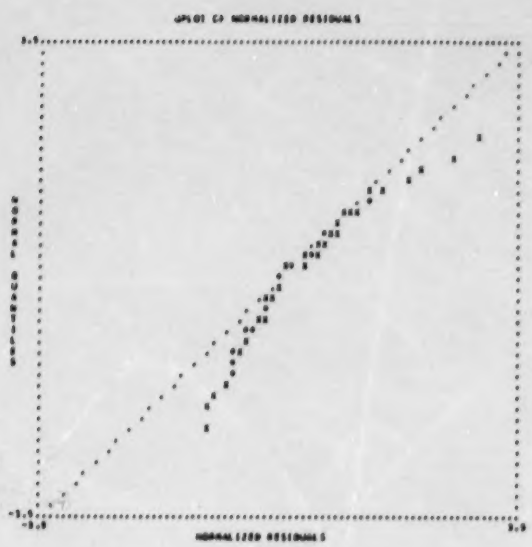
Dropping reading passage items parcels. The results contained in Tables 2 and 3 and Figures 6-8 suggest two conclusions. First, SAT-verbal is not strictly unidimensional and most of the lack of unidimensionality can be attributed to the reading passage items. Second, the content structure for Form V4 differs from that for Forms X2 and Y3. Form V4 needs a second general factor to explain the correlations among the item parcels, a second general factor that Forms X2 and Y3 do not require.

To evaluate the supposition that the reading passage items are the major reason for lack of unidimensionality, factor analyses were conducted on reduced item parcels correlation matrices obtained by excluding the five reading passage items parcels from the matrices. These analyses for the reduced matrices parallel those conducted for the full item parcels correlation matrices.

The data presented in Figures 9-11 parallel that presented in Figures 6-8. Dropping the reading passage items does not result in a drop in the number of first order factors needed to fit the data. The single factor first-order solutions, however, are somewhat better here than they were when the reading passage items parcels were included. Hence, the reading passage items parcels, while a major contributor, are not the sole reason for lack of unidimensionality. Table 4 provides more evidence on this point. From the information presented in this

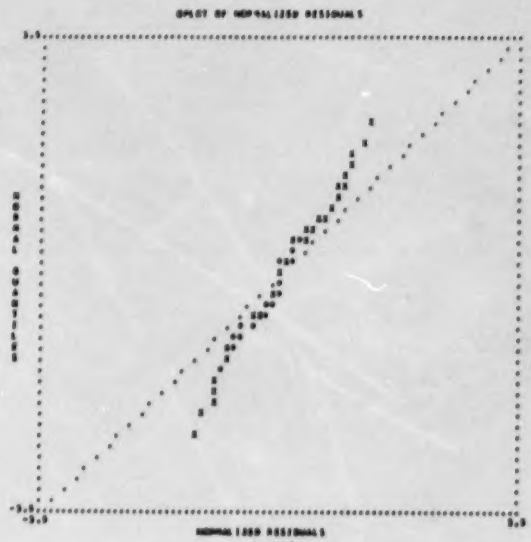
Figure 9

Normalized Residuals Plots and Indices of Fit for SAT-verbal Form V4
(excluding reading passage items parcels)



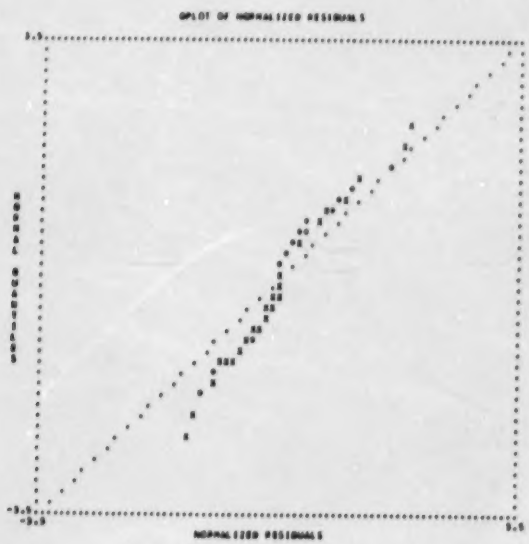
One Factor First Order Solution

Chi Square = 148.90; df = 54
GFI = .986
RMSD = .019



Two Factor First Order Solution

Chi Square = 80.99, df = 41
GFI = .990
RMSD = .013

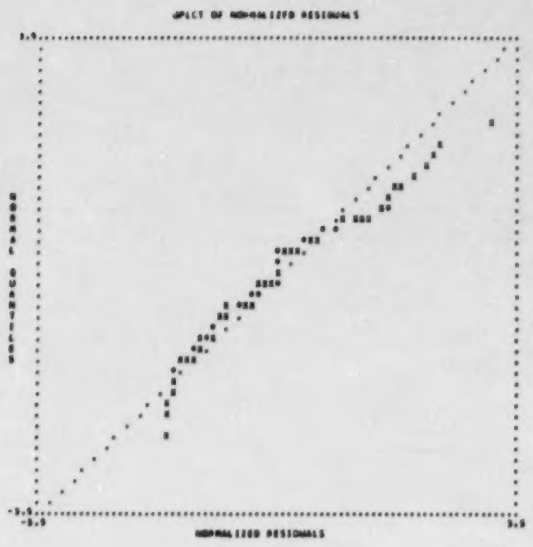


One General Factor and Three Group Factors Solution

Chi Square = 88.92; df = 51
GFI = .991
RMSD = .014

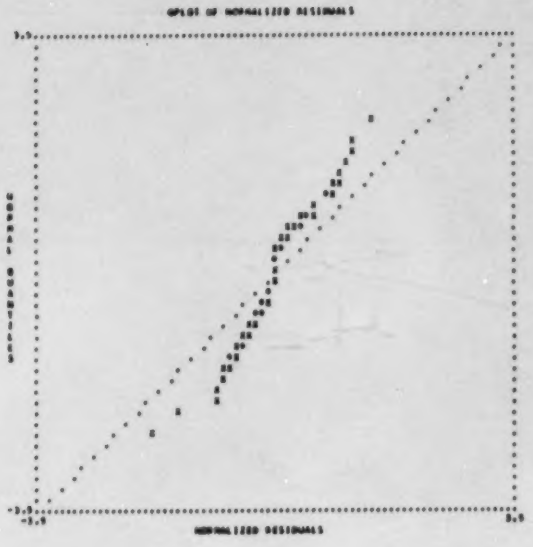
Figure 10

Normalized Residuals Plots and Indices of Fit for SAT-verbal Form X2
(excluding reading passage items parcels)



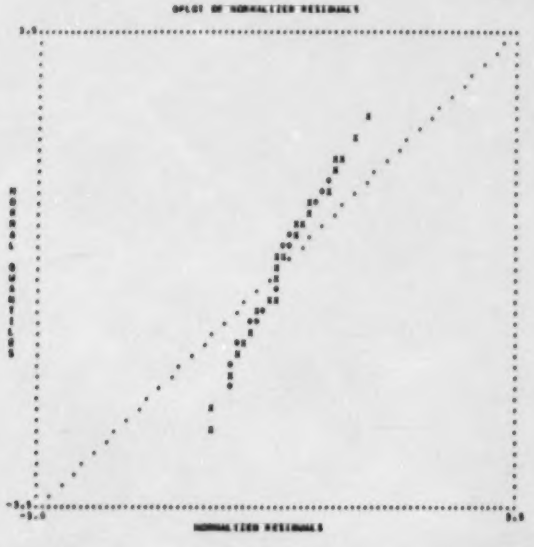
One Factor First Order Solution

Chi Square = 306.31; df = 54
GFI = .967
RMSR = .024



Two Factor First Order Solution

Chi Square = 83.56; df = 41
GFI = .989
RMSR = .012

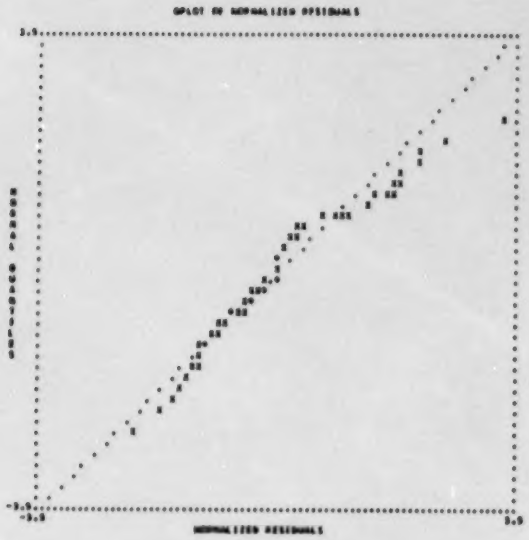


One General Factor and Three Group Factors Solution

Chi Square = 70.59; df = 51
GFI = .993
RMSR = .011

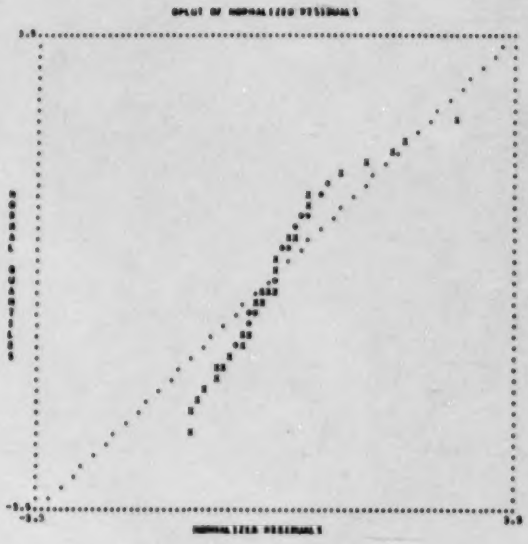
Figure 11

Normalized Residuals Plots and Indices of Fit for SAT-verbal Form Y3
(excluding reading passage items parcels)



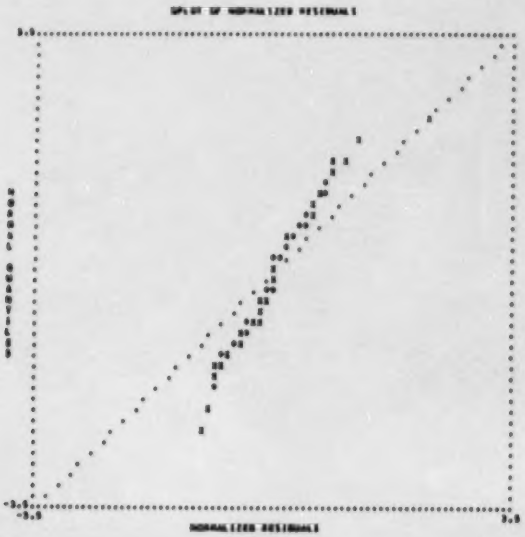
One Factor First Order Solution

Chi Square = 246.28; df = 54
GFI = .975
RMSR = .022



Two Factor First Order Solution

Chi Square = 109.77; df = 41
GFI = .986
RMSR = .014



One General Factor and Three Group Factors Solution

Chi Square = 84.71; df = 51
GFI = .991
RMSR = .013

Table 4

Relative Contributions of One General and Three Group Factors to Variance of First Order Parcel Factors for Three SAT-verbal Forms (excluding reading passage items)

Test Form		First Order Factors			First Order Factor Correlations			
		Sentence Completion I	Antonyms II	Analogies III		I	II	III
V4	general factor	.93	.90	.94	I	1.0		
					II	.92	1.0	
	group factors	.07	.10	.06	III	.93	.92	1.0
						I	II	III
X2	general factor	.96	.92	.82	I	1.0		
					II	.94	1.0	
	group factors	.04[1]	.08	.18	III	.89	.87	1.0
						I	II	III
Y3	general factor	.93	.90	.87	I	1.0		
					II	.92	1.0	
	group factors	.07	.10	.13	III	.90	.88	1.0

[1]Not significantly different from zero ($p<.01$)

table it can be seen that the analogies group factors are sizeable for Form X2 and Y3. One also can see that the structure for Form V4 still gives evidence of being different from that of X2 and Y3. In fact, V4 appears to be the most unidimensional of the three test forms. The structures for X2 and Y3, on the other hand, appear quite parallel. Thus, removing the reading passage items parcels results in data (the remaining item types) that are more unidimensional and clarifies the structural differences between Forms V4 and Forms X2 and Y3.

Mathematics Level II

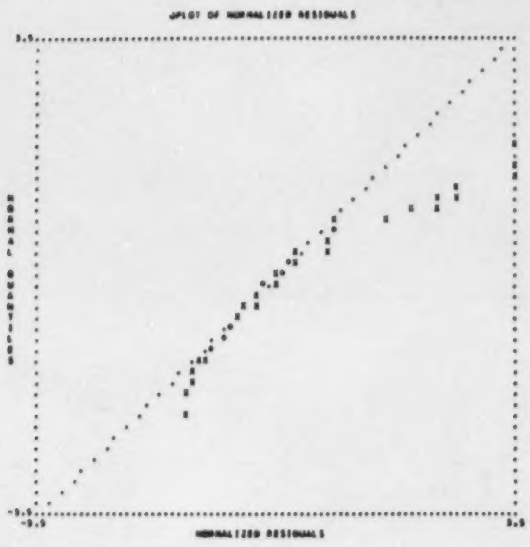
Number of factors. Figure 12 contains plots of normalized residuals and indices of fit for Mathematics Level II Form CC. The top two panels reveal that at least two first-order common factors are needed to fit the Form CC item parcels correlation matrix. Examination of the upper right hand panel reveals that, with the exception of four item parcels correlations, the two common factors provide a reasonable fit to the data. Taking a third common first-order factor (the results are not presented in Figure 12) improves the fit but does not leave many degrees of freedom.

The lower panels of Figure 12 summarize the fit of the restrictive second-order solution of one general factor and four group factors (one for each content area: algebra, geometry, trigonometry, and functions). Using up one less degree of freedom, this second-order solution fits the data very well, indicating that the hypothesized structure for the data is tenable.

Figure 13 contains the summary of indices of fit for Mathematics Level II Form WC. For this test form, two first-order factors provide

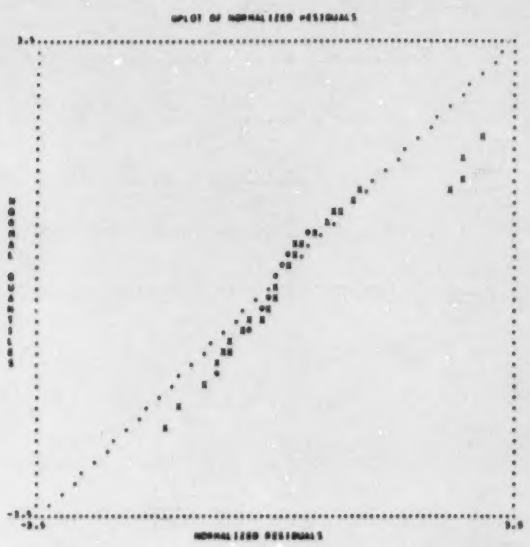
Figure 12

Normalized Residuals Plots and Indices of Fit for Mathematics Level II Form CC



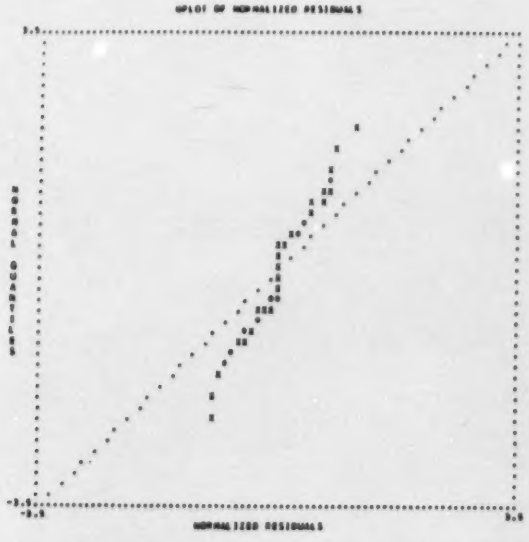
One Factor First Order Solution

Chi Square = 419.08; df = 35
GFI = .937
RMSR = .037



Two Factor First Order Solution

Chi Square = 189.12; df = 24
GFI = .956
RMSR = .023



One General Factor and Four Group Factors Solution

Chi Square = 43.89; df = 25
GFI = .990
RMSR = .012

adequate fit to the data. The second-order solution fits the data even better. The same kind of fit results occur for Mathematics Level II Form AC. These results are summarized in Figure 14.

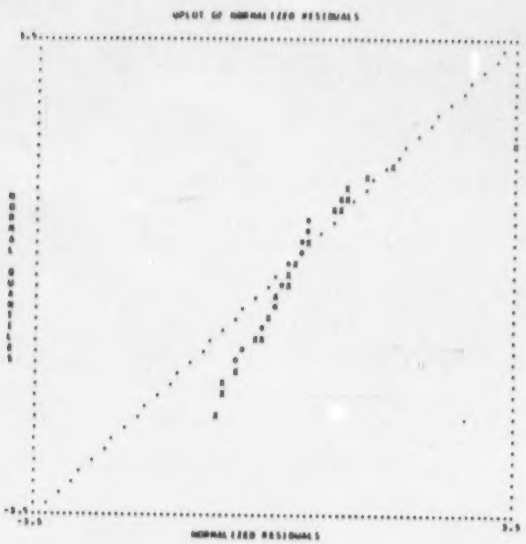
All three forms are fit very well by the second-order solutions of one general factor and four content area group factors. Two common first-order factors are needed to fit the WC and AC item parcels correlation matrices. The CC item parcels correlation matrix, however, is not adequately described by two common first-order factors.

Second-order structures. Table 5 summarizes the contributions of the general and group factors to the first-order factors across all three test forms. As was the case with SAT-verbal, the general factor tends to be large relative to the group factors. On all three forms, the trigonometry factor has the largest group factor, particularly for Forms CC and AC. For Form CC the geometry group factor is quite large; for all forms, the algebra and functions group factors tend to be smallish.

Dropping trigonometry parcels. From the information presented in Table 5, one might infer that the trigonometry item parcels are the primary contributors to lack of unidimensionality. To assess the validity of this inference, factor analyses were conducted on reduced correlation matrices obtained by excluding the two trigonometry parcels for each form. Figures 15-17 summarize the results obtained by fitting various models to the reduced correlation matrices for Forms CC, WC, and AC, respectively. These figures contain information that parallels that found in Figures 12-14. In all three figures, the second-order solution of one general factor and three group factors fits the data very well.

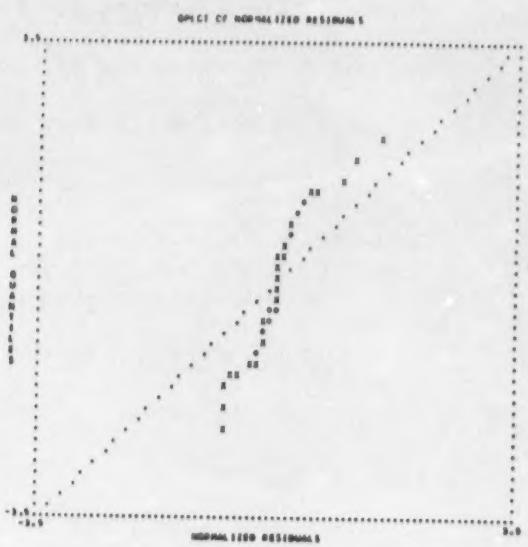
Figure 13

Normalized Residuals Plots and Indices of Fit for Mathematics Level II Form WC



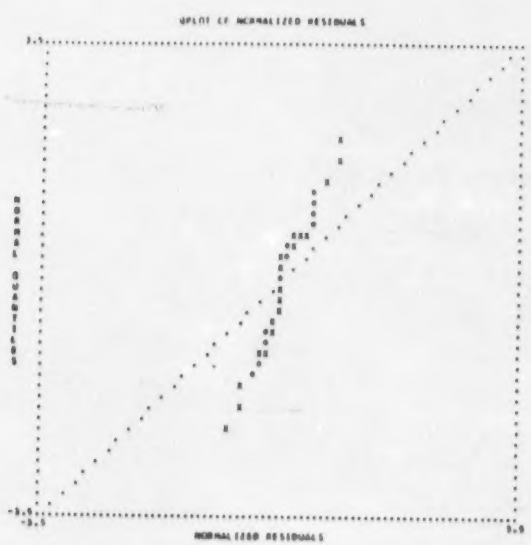
One Factor First Order Solution

Chi Square = 183.06, df = 35
GFI = .969
RMSR = .021



Two Factor First Order Solution

Chi Square = 49; df = 24
GFI = .988
RMSR = .011

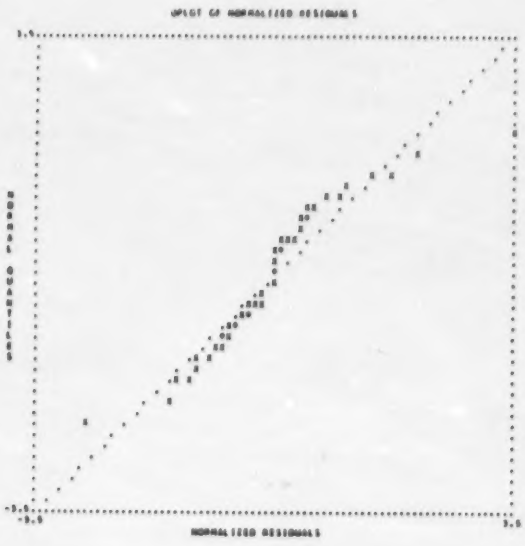


One General Factor and Four Group Factors Solution

Chi Square = 39.68; df = 25
GFI = .991
RMSR = .010

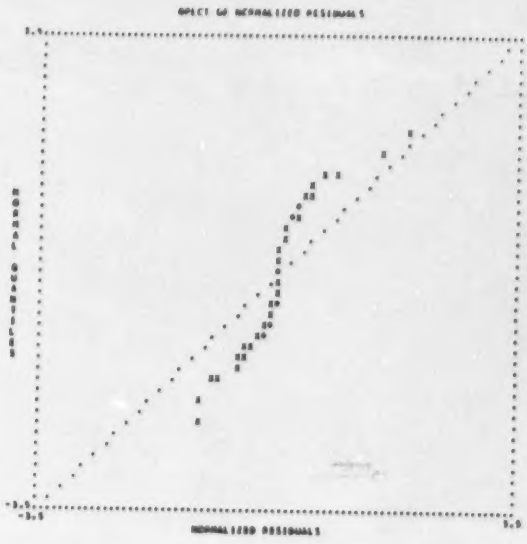
Figure 14

Normalized Residuals Plots and Indices of Fit for Mathematics Level II Form AC



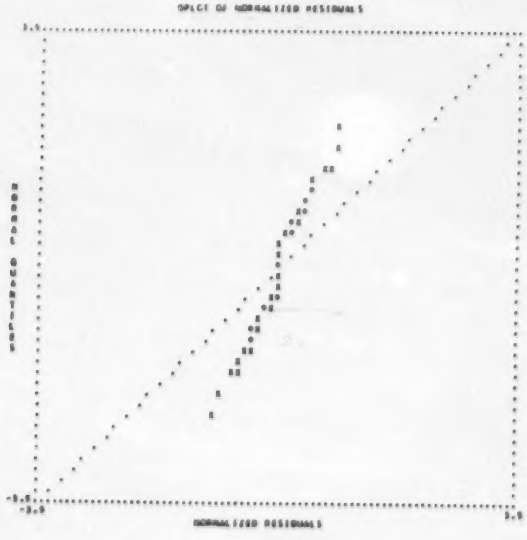
One Factor First Order Solution

Chi Square = 272.69; df = 35
GFI = .963
RMSR = .026



Two Factor First Order Solution

Chi Square = 57.78; df = 24
GFI = .988
RMSR = .011



One General Factor and Four Group Factors Solution

Chi Square = 30.80; df = 25
GFI = .994
RMSR = .008

Table 5

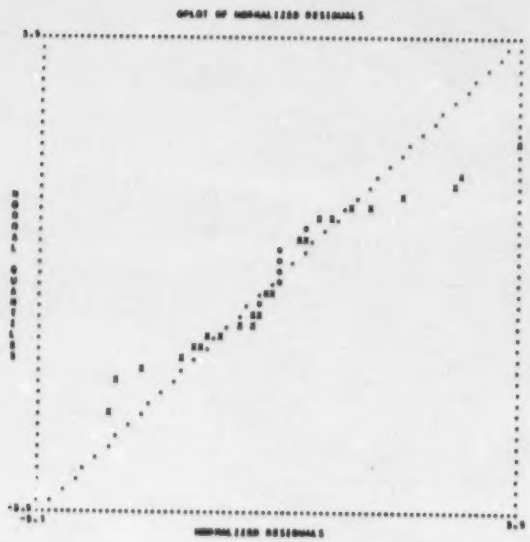
Relative Contributions of One General Factor and Four Group Factors to Variance of First Order Parcel Factors for Three Mathematics Level II Forms

Test Form		First Order Factors				First Order Factor Correlations				
		Algebra I	Geometry II	Trigonometry III	Functions IV					
	general factor	.97	.81	.70	.90	I	1.0			
CC						II	.89	1.0		
						III	.82	.75	1.0	
	group factors	.03[1]	.19	.30	.10	IV	.93	.86	.79	1.0
							I	II	III	IV
	general factor	.93	.88	.82	.94	I	1.0			
WC						II	.91	1.0		
						III	.87	.85	1.0	
	group factors	.07[1]	.12	.18	.06[1]	IV	.94	.91	.88	1.0
							I	II	III	IV
	general factor	.88	1.00	.73	.91	I	1.0			
AC						II	.94	1.0		
						III	.80	.85	1.0	
	group factors	.12	.00[1]	.27	.09	IV	.90	.95	.82	1.0

[1]Not significantly different from zero ($p<.01$)

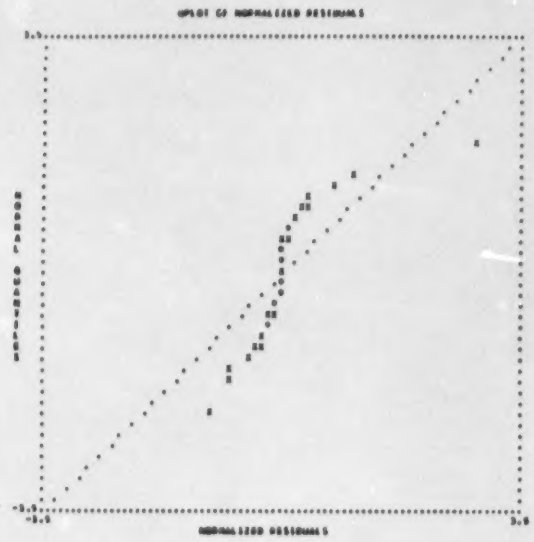
Figure 15

Normalized Residuals Plots and Indices of Fit for Mathematics Level II Form CC
(excluding trigonometry parcels)



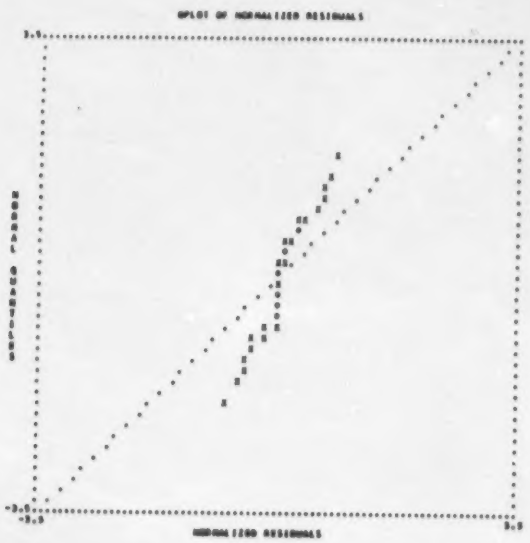
One Factor First Order Solution

Chi Square = 189.46; df = 20
GFI = .956
RMSR = .031



Two Factor First Order Solution

Chi Square = 46.55; df = 11
GFI = .981
RMSR = .015

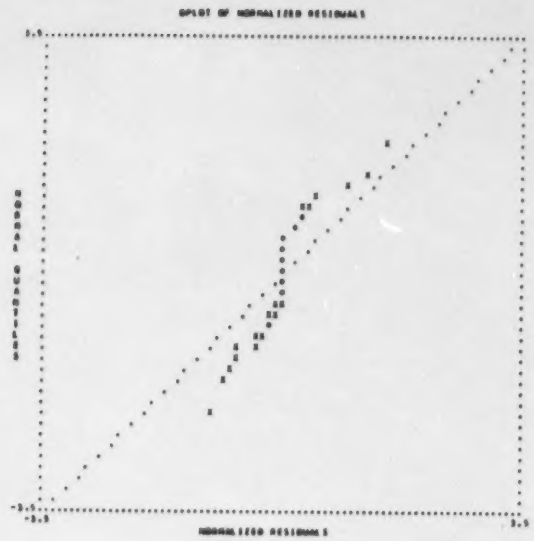


One General Factor and Three Group Factors Solution

Chi Square = 19.74; df = 13
GFI = .993
RMSR = .010

Figure 16

Normalized Residuals Plots and Indices of Fit for Mathematics Level II Form WC (excluding trigonometry parcels)

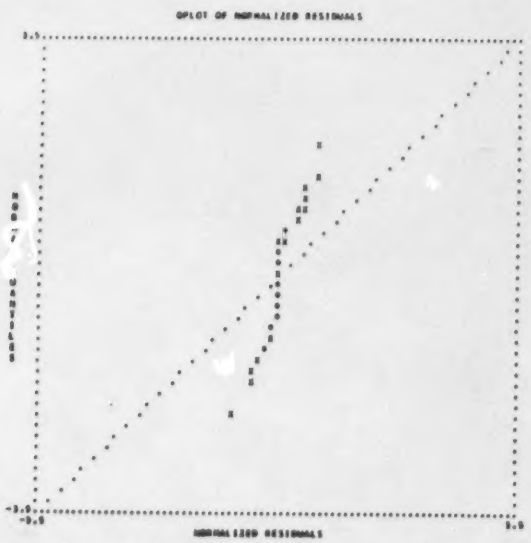


Form WC (excluding trigonometry parcels) basically measured only one general factor. It was impossible to obtain a reasonable two factor solution.

One Factor First Order Solution

Chi Square = 41.88; df = 20
GFI = .990
RMSR = .013

Two Factor First Order Solution

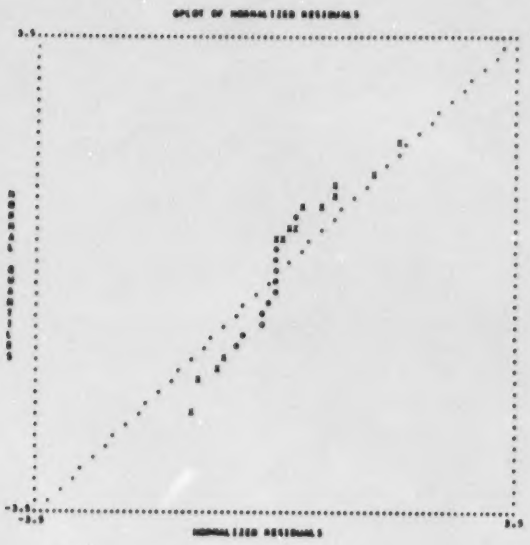


One General Factor and Three Group Factors Solution

Chi Square = 12.14; df = 13
GFI = .996
RMSR = .007

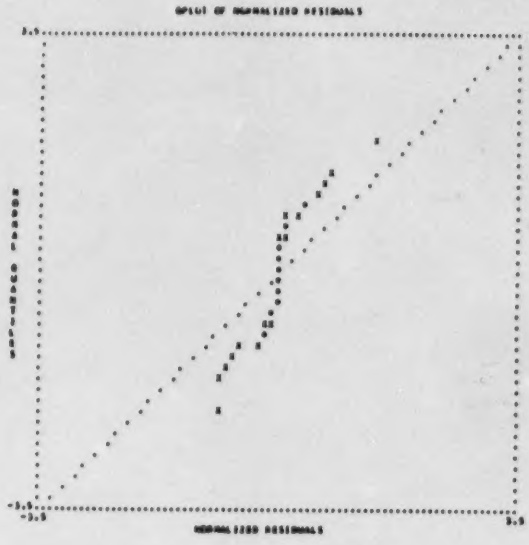
Figure 17

Normalized Residuals Plots and Indices of Fit for Mathematics Level II Form AC
(excluding trigonometry parcels)



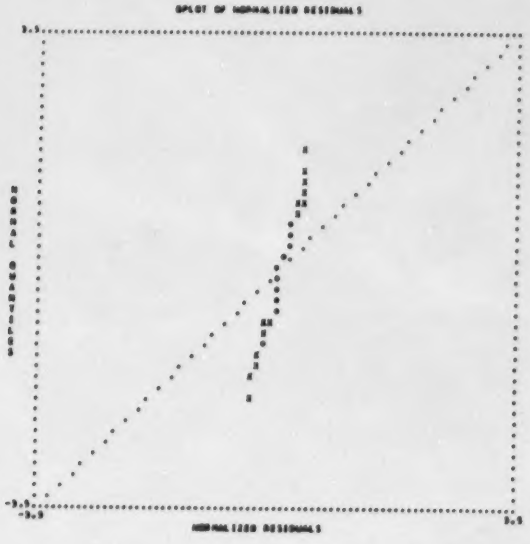
One Factor First Order Solution

Chi Square = 55.07; df = 20
GFI = .989
RMSR = .014



Two Factor First Order Solution

Chi Square = 27.54; df = 11
GFI = .990
RMSR = .010



One General Factor and Three Group Factors Solution

Chi Square = 11.28; df = 13
GFI = .997
RMSR = .006

The most interesting aspect of these figures is the fit of the single first-order common factor solutions, depicted in the upper left panels. For Forms WC and AC, one common factor provides a very tight fit to the reduced correlation matrices. (For Form WC, LISREL V would not even allow a second common factor!) In contrast, Form CC requires a second common first-order factor to achieve a reasonable fit. For Forms WC and AC, removing the trigonometry parcels leaves remaining test items that are very unidimensional. Form CC, however, even after removal of the trigonometry parcels, remains at least two-dimensional.

The unidimensionality of Forms WC and AC that results from excluding the trigonometry parcels is evident from the information presented in Table 6. Note that for these two forms, the first-order correlations are all .91 or higher, and that the contributions of the group factors to first order factor variance are all .10 or less. In contrast, the geometry group factor for Form CC is quite sizeable, while the other two group factors for this form contribute variance that is not significantly ($p<.01$) different from zero. Even after dropping the trigonometry parcels, the structure of Form CC is not unidimensional because of the sizeable geometry group factor.

To summarize, the results of the factor analyses indicate that both the SAT-verbal and the Mathematics Level II forms can be considered to be somewhat multidimensional, and to exhibit some departures from form-to-form parallelism. For SAT-verbal, Form V4 appears to be more unidimensional than the remaining two forms and, as was hypothesized, less parallel to Forms X2 and Y3 than the latter two forms are to each other. Removing the item type for which the group factor contributed

Table 6

Relative Contributions of One General and Three Group Factors to Variance of First Order Parcel Factors for Three Mathematics Level II Forms (excluding trigonometry items)

Test Form		First Order Factors			First Order Factor Correlations			
		Algebra I	Geometry II	Functions III		I	II	III
CC	general factor	.95	.78	.95	I	1.0		
					II	.86	1.0	
	group factors	.05[1]	.22	.05[1]	III	.95	.86	1.0
						I	II	III
WC	general factor	.91	.91	.95	I	1.0		
					II	.91	1.0	
	group factors	.09	.09	.05[1]	III	.93	.93	1.0
						I	II	III
AC	general factor	.90	.97	.93	I	1.0		
					II	.93	1.0	
	group factors	.10	.03[1]	.07[1]	III	.92	.95	1.0

[1]Not significantly different from zero ($p<.01$)

the most to parcel variance (reading passage items), although providing data of a more unidimensional nature, did not result in what could be considered a truly unidimensional set of items for any of the test forms. Of the Mathematics Level II forms investigated, Form CC appeared to be less unidimensional than Forms AC and WC. Form CC also appeared to be less parallel to Forms WC and AC than these two forms were to each other. Removal of the content category (trigonometry) that contained item parcels for which the group factor contributed most to parcel variance did not result in unidimensionality for the remaining items in Form CC. However, removal of item parcels in this content area did result in virtually unidimensional data for the remaining items in Forms WC and AC.

DISCUSSION

This research was conducted in an attempt to develop a better understanding of the relationship between violations of the assumption of unidimensionality and the quality of IRT equating results. Examination of this relationship is hampered by the difficulties associated with assessing dimensionality when using binary item data. In an attempt to circumvent some of these difficulties, item parcels were constructed. Construction of these parcels was guided by content and item type considerations, and a desire to produce correlations that could be fit by linear factor models. The resultant correlation matrices were subjected to a series of confirmatory factor analyses employing the LISREL V model.

This series of analyses did provide a better understanding of the relationship between violations of the assumption of unidimensionality

and the quality of IRT equatings. For example, the Mathematics Level II equating results were viewed as superior to the SAT-verbal equating results, and the dimensionality analyses revealed that the Mathematics Level II item parcels were more nearly unidimensional than the SAT-verbal item parcels. In addition, the dimensionality analyses verifed that SAT-verbal Form V4 and Mathematics Level II Form CC were each less parallel to the other two forms in their respective equating chains than the other forms (SAT-verbal X2 and Y3 and Mathematics Level II AC and WC) were to each other.

While the research presented in this paper has provided a better understanding of the relationship between the assumption of unidimensionality and the quality of IRT equating, there is definitely soom room for enhancement. Refinements of the methodology for assessing dimensionality that was used in this study are needed. For example, conducting a series of dimensionality analyses throughout the entire equating chain (for each of the tests studied) should improve understanding, particularly if item parcels containing the common (equating) items appeared in adjacent analyses. Use of common item parcels in adjacent analyses would make analyses of variance-covariance matrices (instead of correlations) more meaningful, provided that item parcel construction could be refined to produce parcels with approximately equal variances as well as equal means. Given the strict adherence to item type composition observed for the Scholastic Aptitude Test, the verbal and mathematical sections of this test seem most amenable to a more thorough dimensionality analysis. This more thorough analysis should uncover more general (and perhaps contrasting) trends in

dimensionality and form-to-form parallelism that could be related to the quality of IRT equating. Eventually, this approach might yield diagnostics that could be used to arrive at more informed equating decisions.

In the interim, it is reassuring to note that, despite some variation in form-to-form parallelism and some departures from unidimensionality, both the SAT-verbal and Mathematics Level II IRT equating results were quite reasonable. Perhaps, as Divgi (1981b) might argue, IRT equating is robust to violations of unidimensionality because test scores are involved, not predictions of individual item responses. Or, as Drasgow and Parsons (in press) might argue, IRT equating works when the general factor is prepotent, i.e., accounts for much of the variance in the data. (In this study, the general factors in the SAT-verbal and Mathematics Level II analyses were very large.) Further dimensionality assessment studies should provide more answers, generate more questions, and ultimately lead to improved empirical techniques for dimensionality assessment as well as a firmer conceptual framework for evaluating IRT equatings.

References

Bejar, I. I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 1980, 17, 283-296.

Bock, R. D., and Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Bock, R. D., and Aitken, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Cattell, R. B., and Burdsal, C. A. The radial parcel double factoring design: A solution to the item-vs-parcel controversy. Multivariate Behavioral Research, 1975, 10, 165-179.

Cook, L. L., and Eignor, D. R. Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.) Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.(a)

Cook, L. L., and Eignor, D. R. An investigation of the feasibility of applying item response theory to equate achievement tests. Paper presented at the annual meeting of AERA, Montreal, 1983. (b)

Divgi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981.(a)

Divgi, D. R. Potential pitfalls in applications of item response theory. Paper presented at the annual meeting of NCME, Los Angeles, 1981.(b)

Drasgow, F., and Parsons, C. K. Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 1983, in press.

Fischer, G. H. Probabilistic test models and their applications. German Journal of Psychology, 1978, 2, 298-319.

Gustafsson, J-E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980, 33, 205-233.

Hambleton, R. K. (Ed.) Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R. K., and Rovinelli, R. J. Assessing the dimensionality of a set of test items. Paper presented at the annual meeting of AERA, Montreal, 1983.

Hattie, J. A. Decision criteria for determining unidimensionality. Unpublished doctoral dissertation, University of Toronto, 1981.

Hecht, L. W., and Swineford, F. Item analysis at Educational Testing Service. Princeton, NJ: Educational Testing Service, 1981.

Joreskog, K. G., and Sorbom, D. LISREL V-Analysis of linear structural relationships by the method of maximum likelihood. Chicago, IL: International Educational Services, 1981.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison Wesley, 1968.

McDonald, R. P. Nonlinear factor analysis. Psychometric Monographs, 1967, no.15.

McDonald, R. P. The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 1981, 34, 100-117.

McDonald, R. P. Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 1982, 6, 379-396.

McDonald, R. P., and Ahlawat, K. S. Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 1974, 27, 82-99.

Petersen, N. S., Cook, L. L., and Marco, G. L. Using item response theory to equate Scholastic Aptitude Test scores. Paper presented at the annual meeting of APA, Washington, 1982.

Petersen, N. S., Cook, L. L., and Stocking, M. L. Scale drift: A comparative study of IRT versus linear equating methods. Journal of Educational Statistics, in press.

Schmid, J., and Leiman, J. The development of hierarchical factor solutions. Psychometrika, 1957, 22, 53-61.

Swinton, S. S., and Powers, D. E. A factor analytic study of the restructured GRE Aptitude Test. GRE Board Professional Report GREB No. 77-6P. Princeton, NJ: Educational Testing Service, 1980.

Van den Wollenberg, A. L. Two new test statistics for the Rasch model. Psychometrika, 1982, 47, 123-140.(a)

Van den Wollenberg, A. L. A simple and effective method to test the dimensionality axiom of the Rasch model. Applied Psychological Measurement, 1982, 6, 83-91.(b)

Wingersky, M. S., Barton, M. A., and Lord, F. M. LOGIST V user's guide. Princeton, NJ: Educational Testing Service, 1982.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.